Verification of Drosophila Sequence Assembly

using restriction digest BAC fingerprints derived from multiple enzymes

Krzywinski M, Schein J, Chiu R, Bosdet I, Mathewson C, Wye N, Barber S, Brown-John M, Chand S Cloutier A, Masson A, Mayo M, Olson T, Jones S, Hoskins R†, Celniker S†, Rubin G‡, Marra M

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

Canada's Michael Smith
Genome Sciences Centre
www.bcgsc.ca

1. Abstract

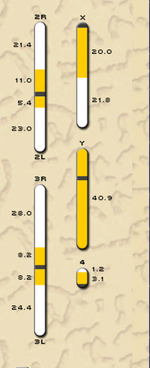
The annotated *D. melanogaster* genomic sequence is currently in its third revision (Release 3) and covers nearly all of the 120 Mb euchromatic DNA. The sequence assembly is curated by the Berkeley Drosophila Genome Project (BDGP) and comprised of data produced at Celera Genomics, Genoscope, Lawrence-Berkeley National Labs, and the Human Genome Sequencing Center at the Baylor College of Medicine. Drosophila continues to play a major role in providing a model for inheritance and gene interaction and a high quality assembly is required to ensure accuracy of sequence-based analysis. To this end, we have developed an automated data analysis and visualization pipeline for verification of the sequence assembly using multiple restriction enzyme digests of tiling path BAC clones. Various types of repeat regions produce incorrect, but self-consistent, sequence assemblies. These errors are very difficult to spot without an external validation method. The fingerprint verification method offers several benefits: the sequence is verified by an independent laboratory process, the fingerprints are robust in elucidating repeat elements and the data processing pipeline is extensible and can be adapted to any sequence data.

A set of 988 tiling path clones spanning the euchromatic portion of the genome were selected. Each clone was independently fingerprinted using 5 different restriction enzymes. The enzymes were chosen to maximize coverage of the sequence with fragments in the size range of 1-20 kb to facilitate accurate sizing. The enzymes selected were ApaLI, BamHI, EcoRI, HindIII and XhoI. This combination provides coverage by at least two, three and four optimally-sized fragments for 99.9%, 98% and 87% of the sequence, respectively.

An *in silico* fingerprint of each clone was derived from the sequence and compared to its experimental counterpart using a Needleman-Wunsch alignment and a 2% fragment size tolerance. Each base of the sequence is assigned a verification depth that corresponds to the number of experimentally verified *in silico* fragments containing that sequence location. The average verification depth is used as a measure of overall verification. We have devised various figures of merit to identify clones with unverified subsequences and to categorize the discrepancies. An interactive web-based system has been created to visualize verification coverage.

To date, we have discovered 105 BACs (12% of 891 active BACs in our validation pipeline, see Figure 10) whose validation profile indicates a possible sequence assembly error in regions totalling 709kb. In 29 of these cases (3.3%, seven are Phase 2), within regions spanning 274kb, the inconsistency was verified and the assembly construction in these regions will be thoroughly rechecked. The other 76 BACs require additional digests to validate the sequence assembly because the regions of inconsistencies are represented by bands that are too large (> 30kb) or are too numerous (>3X copy number). The method described here will also be applied to verification of heterochromatic DNA sequence, which is being generated using smaller clones. We anticipate that this fingerprint-based sequence verification methodology can positively impact the final sequence assembly quality of other organisms such as human, mouse and rat.

Drosophila Chromosomes



Assembly History

Release 1

March 2000

Celera provides 12X coverage. BDGP provides BAC physical map, 26 Mb finished sequence and 1.5X shotgun for each tiling set BAC

Release 2

October 2000

Celera/BDGP fill 330 gaps, leaving 1300 remaining gaps in the assembly

Release 3

July 2002

2L,2R,3R,4 and 12.20X finished by BDGP, 3L and 1-11X finished by HGSC

Verification Status

by BACs

988 BACs in analysis pipeline (891 active BACs, see Figure 10).

76 BACs span possible sequence assembly errors (SAEs) (11%)

29 BACs span verified SAEs (3.3%)

786 BACs are validated (no 0-depth regions)

485 BACs are entirely validated by at least 2 enzymes (no 1-depth regions)

by Sequence

116.5Mb covered by active BACs

709kb in 0,1-depth regions with possible SAEs

274kb in 0,1-depth regions for BACs with verified SAEs

2L: 3 kb (1 BAC)

2R: 11 kb (2 BACs)

3L: 90 kb (7 BACs)

3R: 35 kb (5 BACs)

4: 11 kb (1 BAC)

X: 124 kb (13 BACs)

2. Validation Methodology

The sequence spanned by each tiling set BAC is validated by comparing its experimental and *in silico* fingerprints derived from digests using 5 different restriction enzymes. The specific combination of restriction enzymes was chosen to maximize the depth of validation by accurately sized fragments and minimize the effect that undetectable or unsizeable fragments have on the validation (Figure 1). Fragments which are <600bp or >30kb are not reliably sized by our agarose electrophoresis method. The largest marker fragment is 29,950bp and fragments larger than this cannot be accurately sized. Fragments smaller than 600bp are very diffuse and are not always detected. The enzyme combination was therefore chosen to maximize coverage by fragments in the range 1-20kb.

In Figure 1, the quantity $S(i)$ corresponds to the fraction of genomic sequence in which every base pair is covered by at least i optimally sized fragments for a given enzyme combination.

The 5-enzyme combination was selected from about 200,000 computationally simulated combinations. The combinations were scored on the merits of ease of use and $S(i)$ values. Some enzymes which contributed to better coverage were either not available in high concentration or require specialized laboratory conditions for reproducible digests.

The best practical choice of enzymes which yield consistent high-quality fingerprints and provide optimum coverage for Drosophila sequence is **ApaLI** (g.tgcac), **BamHI** (g.gatcc), **EcoRI** (g.aattc), **HindIII** (a.agctt) and **XhoI** (c.tcgac). The average cut site GC content is 53% (vs 46% for the Drosophila genome).

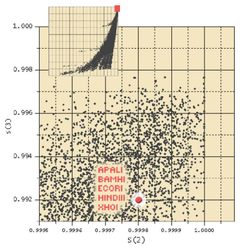


Figure 1 Profile of S(3) vs S(2) for all simulated combinations (*inset*). The extent of coverage by our 5 enzyme combination is highlighted in the zoomed part of the plot.

The *in silico* and experimental fingerprints were compared using a global alignment algorithm with a uniform, relative size tolerance of 2%. The alignment attempts to match as many fragments as possible that are within 2% of their size between the fingerprints while minimizing the sum of differences for all matched fragments. Because the orientation of the insert relative to the vector in the BAC is not known (Figure 2), two *in silico* fingerprints are generated. The orientation is deduced from the *in silico* fingerprint which was the better match to the experimental fingerprint.

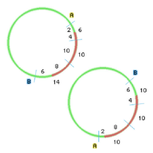


Figure 2 BAC insert can be incorporated into its vector in one of two orientations.

For a sequence region, the number of experimental digests which validate the size of the region's *in silico* fragments is called the **validation depth** and is used as a metric of validation. Regions with 0-depth validation regions are considered unvalidated, and span possible sequence assembly errors. Areas with 1-depth validation typically correspond to sequence regions with a sparse restriction cut site distribution, which may require additional digests to be validated.

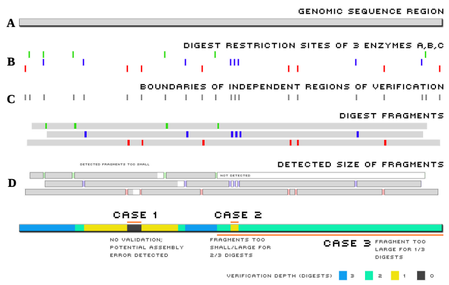


Figure 3 Illustration of the fingerprint validation methodology. A genomic sequence region (A) is digested independently with multiple enzymes (B). The density of restriction sites (C), found by performing an *in silico* digest of the corresponding assembled sequence, determines the resolution of the method. Detected fragments are used to validate regions of the sequence. Undetected fragments or fragments detected to be inconsistently sized indicate possible error in the sequence assembly (D). The number of fingerprints verifying a given region of sequence, is used to quantify the degree of validation and is called the **validation depth**.

3. Verification Process

The verification process is based on the concept of n-depth validation described in the previous section. Any BAC with N experimental fingerprints may have up to N-depth validation for its regions. Typically, no BAC has N-depth validation everywhere. Instead, a distribution is seen because a number of fragments for some digests are not detected or are not detectable. The average validation depth is found to be 0.82 per experimental fingerprint. Any BACs with regions of 0/1-depth validation are examined, since they may span possible sequence assembly errors.

The verification process depends on high-quality and high-resolution fingerprints. Figure 4 shows the experimental error for sizing fingerprint fragments.

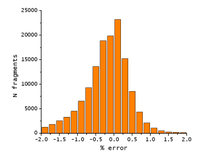


Figure 4 Experimental and *in silico* fingerprint fragments are matched using a 2% size tolerance and a global alignment algorithm. Out of 137,000 matched fragments the sizing error was found to be 0.2±0.6%.

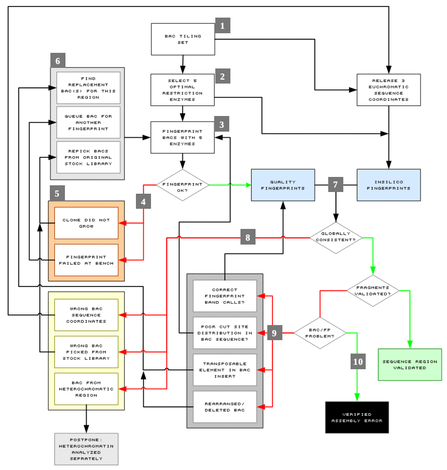


Figure 5 Outline of the data and process flow. Using an initial tiling set of BACs (1), optimal enzymes are selected (2) and all BACs are fingerprinted with each enzyme in independent digests (3). Any fingerprints which fail (4) are categorized (5) and appropriate action is taken to correct the problem and ensure that an acceptable fingerprint is obtained or alternative BAC is chosen (6). Acceptable quality fingerprints are compared (7) to *in silico* fingerprints to derive an overall validation score for the BAC. Comparisons with global inconsistencies are typically due to BACs which contain E coli elements, are misidentified or are heterochromatic (8). Unvalidated fragments which can be explained by a local problem (9) are distinguished from those where the lack of validation appears to be genuine (10). The latter case is identified as a verified sequence assembly error and the assembly in this region is scrutinized.

4. REDEYE – An Online Validation Tool And Viewer

To track individual BACs through the validation process (Figure 5) and facilitate communication throughout the project, we have developed a web-based verification tool. Written in Perl/Mason, Redeye is an interactive visualization tool which provides an environment providing the following functionality:

- BAC digest maps
- individual fingerprints
- fingerprint fragment accounting
- summary coverage and validation statistics
- overall BAC validation
- overall chromosome validation and coverage
- prioritized BAC status/action annotation
- user authentication and tracking

Figure 6 User login screen in Redeye. The name of the application was inspired by fly photos taken by Sharon Gorski (GSC) using a dissection microscope.

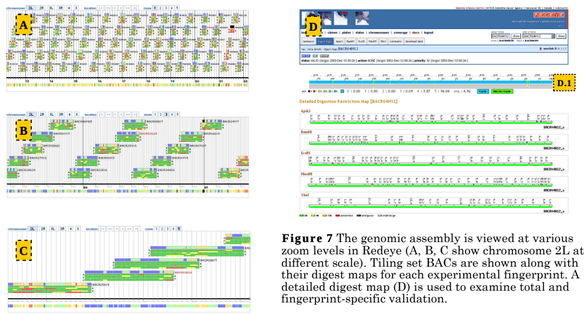


Figure 7 The genomic assembly is viewed at various zoom levels in Redeye (A, B, C show chromosome 2L at different scale). Tiling set BACs are shown along with their digest maps for each experimental fingerprint. A detailed digest map (D) is used to examine total and fingerprint-specific validation.

To examine the validation results of a BAC sequence, a user would typically start by looking at the digest maps for the BAC (Figure 7). Fragments in each fingerprint are coloured by the difference in size between the observed fragment and *in silico* fragment (categories exist for 2%, 4%, 10%, or >10% error, as well as ambiguous matches). Fragment pairs with an error of <2% are considered significant matches and all others are considered as non-matches. The fingerprints are stacked on the same sequence length scale and any overlapping regions of poor validation (see next section) may indicate a sequence assembly error. The overall validation depth for any region, mapped by a coloured bar on top of the digest maps (Figure D, section D.1), is the number of matched fragments in all fingerprints which overlap the region.

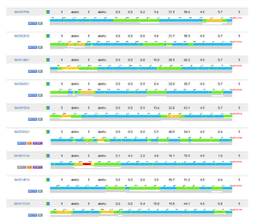


Figure 8 Summary validation depth and status annotations for various BACs.

5. BAC Validation

In this section, different ways of viewing a BAC analysis are shown, using BACR10M16 as an example. This clone is assigned a SEQERR status because the BAC has a high average validation depth (4.03) but also has regions of 0-depth validation (2.74%). Since no inconsistencies were found, the sequence error was verified and the assembly in the region of this BAC will be examined.

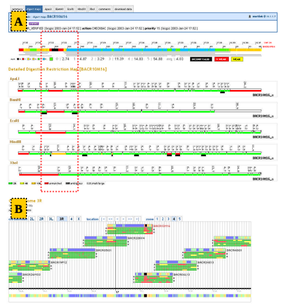


Figure 8 *panel A* | Detailed digest maps for BACR10M16 show a region with 0-depth validation. *panel B* | View of local assembly region for BACR10M16. The indication of a potential sequence assembly error is strengthened by the fact that BACR06L13 overlaps with BACR10M16 and shows the same pattern of 0-depth validation for the same region of the assembly.

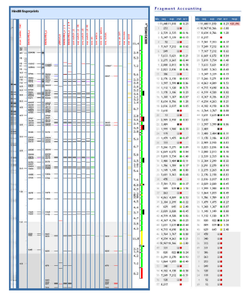


Figure 9 To verify the fingerprint fragment calls and analyze the discrepancy between experimental and *in silico* fingerprints, a detailed fingerprint view is used. All fragments are shown in a table for easy reference. Any fragments <600bp or >30,000bp are outside the reliable sizing range and it is not expected that they would be validated.

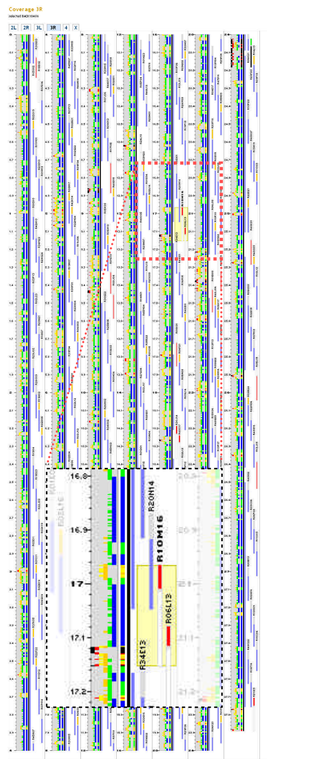


Figure 10 BACR10M16 is located on chromosome 3R. Shown here is the entire chromosome with validation depth shown for each region. The BAC in question is selected and a zoomed view of the region covered by the BAC is shown. This view allows a user to rapidly identify regions of interest.

5. Current Sequence Validation Status

We are currently analyzing discrepancies between the experimental and *in silico* fingerprints, performing additional fingerprints and extending the analysis to cover the heterochromatic portion of the assembly. **We have identified 105 BACs (11.8%) with unvalidated regions, with 29 (3.3%) of these believed to represent authentic assembly errors.** 786 BACs (88%) have no unvalidated regions but 299 (34%) have some regions of 1-depth validation. This group of 299 BACs is being analyzed further to determine the nature of the discrepancy and the appropriate course of action to resolve it.

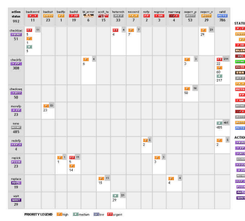
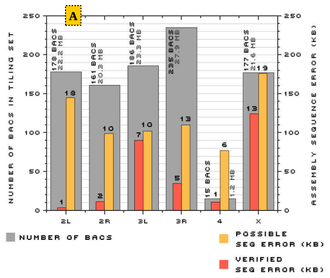


Figure 10 The breakdown of the number of clones at various states of analysis. The last three columns represent BACs without any global inconsistencies (see Figure 5). BACs are further stratified into an action and priority category.



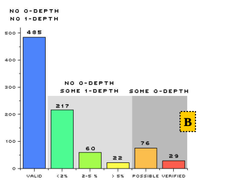


Figure 11 *panel A (top)* | The breakdown of the number of BACs on each chromosome, the size of the chromosome and the number of unvalidated BACs and size of unvalidated regions. For example, on 3R we have analyzed 235 BACs and to date found 5/235 BACs with unvalidated regions spanning 35kb. An additional 13/235 BACs have potential errors spanning 110 kb. *panel B (left)* | The number of validated BACs, poorly validated (require additional analysis) and BACs with unvalidated regions.

Acknowledgements

Clone libraries | RPCI-98/CH-221/CH-223 BAC PAC Resource Centre www.chori.org/bacpac | BAC(N,H) MRC Geneservice www.hgmp.mrc.ac.uk/geneservice | *BAC physical map and sequence* | Berkeley Drosophila Genome Project www.fruitfly.org

†Berkeley Drosophila Genome Project, Lawrence Berkeley National Laboratory, Berkeley, CA, USA | ‡Dept of Molecular and Cellular Biology and Howard Hughes Medical Institute, UC Berkeley, Berkeley, CA, USA